
Seyed M. M. Rizi, Maciej M. Latek, and Armando Geller

Center for Social Complexity
George Mason University
4400 University Drive
Fairfax VA 22030, USA
{smussavi,mlatek,ageller1}@gmu.edu

Abstract. Links between licit and illicit economies fuel processes of conflict in countries mired in irregular warfare. However, it is often infeasible to gather demographic and economic data in the war zone that illuminate such links. Combining remote sensing data with local surveys offers an alternative to direct data collection. We present and validate an implementation of such an approach in order to recover an agentized household representation of the rural population in Afghanistan where poppy cultivation and opium trafficking have influenced the course of the current conflict significantly.

Key words: Multiagent Simulation, Remote Sensing, Bounded Rationality, Irregular Warfare, Empirical Models

1 Introduction

Understanding irregular warfare requires examining the links between licit and illicit economic activities that finance the warring parties’ operations and sustain conflict-torn populations. Poppy cultivation and opium trafficking are the most conspicuous illicit economic activities that have since 2001 influenced the course of the Afghan conflict [1]. Although detailed field studies show that security conditions and alternative economic opportunities drive the Afghan farmers’ decision to cultivate poppy [2, 3], multiagent models of drugs supply chain [4] and macroeconomic structural equation models of drug exports [1] have failed to incorporate the objectives that players in the Afghan drug industry pursue and the biophysical, geographic and resource constraints they face. This shortcoming has diminished the usefulness of [1, 4] and similar work as policy analysis and support tools.

* Funding for this work was provided in part by the Center for Social Complexity at George Mason University and by the Office of Naval Research (ONR) grant No N00014–08–1–0378. Opinions expressed herein are solely those of the authors, not of George Mason University or ONR.
Fig. 1. UML class diagram of the static structure of the Afghan agrarian economy model based on entities and aggregations recovered from data provided by the U.S. National Geospatial Intelligence Agency (NGIA), the U.S. Geological Survey (USGS), the United Nations Office on Drugs and Crime (UNODC), the United Nations Food and Agriculture Organization (FAO), Afghanistan Central Statistics Organization (CSO), Afghanistan Information Management Services (AIMS) and Oak Ridge National Labs (ORNL).

Some recent multiagent models have begun to use fine-grained georeferenced data [5,6]. We contribute to this research by instantiating a country-wide multiagent model of the drug industry in Afghanistan with purposive and boundedly rational agents that is empirically informed by the joint spatial distribution of farmers and land ownership on the Afghan landscape. We combine the following open source, partial data that contain only subsets of dimensions required to instantiate a multiagent model, to turn players in the Afghan drug industry into populations of heterogeneous agents with realistic material attributes such as land ownership, wealth and location in order to build a multiagent model of the Afghan drug industry that includes poppy cultivation, and opium trade and trafficking:

(a) Remote sensing data on agricultural production and rural population in Afghanistan [7,8];

(b) Spatial panel data for the quantity and prices in various drug, food and produce markets in Afghanistan [1,9];

(c) Spatial panel data on security incidents in Afghanistan [10] and georeferenced Afghan farmer surveys [11].

Coded in MASON [12], the model represents biophysical and climate data layers explicitly and houses heterogeneous populations of $1.5 \times 10^6$ farmer households, $3 \times 10^5$ agricultural traders, $10^5$ drug traffickers and 50 major drug traffickers. These classes of agents interact with one another and with the urban population...
in regional markets monitored by Afghan insurgents, NATO and a medley of security forces such as the Afghan National Army, Afghan National Police and Afghan Counternarcotics Police. Featuring a variety of plausible agent responses to counternarcotics policies such as crop eradication and drug interdiction, the model yields district-level predictions of the real, licit and illicit agricultural output, and flows of bribes and taxes to the government and funds to insurgents.

This paper describes procedures that we used to merge remote sensing data and population surveys in order to derive the static structure and states of farmer households and recover their attributes such as patterns of land ownership and wealth. We present data sources, processing steps and a validation exercise for the recovered structure.

2 Input Data

As shown in Fig. 2, country-scale marginal and joint small-sample distributions of farmer and land attributes are available from the following data sources:

(1) AIMS [8] provides us with
   (a) settlement locations without population information produced by the NGIA on Fig. 2(a);
   (b) available land types and existing irrigation mechanisms for land plots without links to the farmer households that own land on Fig. 2(c);
   (c) road networks on Fig. 2(d).

(2) CSO [13] contains information on
   (a) population of cities with more than 3000 inhabitants on Fig. 2(d);
   (b) district-level rural and urban populations on Fig. 2(b).

(3) Fig. 2(e) presents locations of surveyed farmer households, colored by the total wealth of the village they belong to, in a country-wide 2002 FAO survey of 5000 farmer households with 10–20 households per surveyed village [11]. Fig. 3(a) presents the joint distribution of household size and the wealth of surveyed farms, expressed in jeribs, a traditional unit of land in Afghanistan, standardized to 0.2 hectare.

3 Processing

To create a virtual rural population for Afghanistan from the available partial data we proceed as follows:

(1) For each district, subtract the urban population from the total population count recorded in the CSO layer;

(2) Assign all arable land from the AIMS layer to the nearest village from the NGIA layer. The distance used to determine the nearest settlement for a given land lot is based on the USGS digital elevation travel-time metric that corrects geographical distance for land elevation and the presence of roads [14].
Fig. 2. Vector data layers provided by AIMS, COS, ORNL and FAO. Population counts on Fig. 2(b) come from CSO estimates for 2000 and were overlaid on the AIMS district border layer. The bottom row presents a 1-km\textsuperscript{2} population raster by ORNL for 2008 and a georeferenced survey of 5000 farmers conducted in 2002–2003 by the FAO. The ORNL population dataset is retained for validating our procedures, not to create farmer households.
This loop yields a set of villages, each with some lots of land in any of the five land-type categories shown in Table 1. The next loop repeats the following algorithm for each district:

(3) While there is unassigned population in the district:
   (a) Pick a village with probability proportional to the unassigned land in the village, using the weights in Table 1 to scalarize the acreage of different land types.
   (b) Pick a random surveyed farmer household from [11] with a probability proportional to the weighted sum of (a) the geographic distance to the chosen village and (b) the difference between the total wealth of the village a given surveyed farmer belongs to and the chosen village;
   (c) Create a virtual household with the same attributes as the surveyed household and place it in the chosen village;
   (d) Assign the required number of land lots to the chosen household, unless land has been already exhausted;
   (e) Recalculate unassigned district population and unassigned village land wealth.

<table>
<thead>
<tr>
<th>Land type</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigated, extensively cultivated</td>
<td>3</td>
</tr>
<tr>
<td>Irrigated, intensively cultivated</td>
<td>1.5</td>
</tr>
<tr>
<td>Irrigated, intermittently cultivated</td>
<td>1</td>
</tr>
<tr>
<td>Rain-fed, flat-lying areas</td>
<td>0.75</td>
</tr>
<tr>
<td>Rain-fed, sloped areas</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 1. Five prevalent land types in Afghanistan categorized by water source and cultivation intensity along with wealth weights assigned to each type. The weights denote relative annual wheat yield of each type of land.

This preferential attachment procedure through which a village is more likely to receive a new household if it has a higher percentage of unassigned land, creates a population of farmer households, each of which assigned a particular size and a set of land lots while preserving the following statistics that guarantee a reasonable degree of fidelity to the actual data generating processes:

(a) District level population counts;
(b) Regional availability of land and land types;
(c) Joint distribution of household size and wealth, conditioned on village wealth.

4 Outputs and Validation

In order to validate the algorithm described in Section 3, we compare the distribution of the simulated population with that in the ORNL data [7] that
contains multi-sourced worldwide population data compiled on a 1 km$^2$ grid. We can compare the [7] extract for Afghanistan with our simulated population by Spearman correlation $\rho_h$ between the two data layers at each spatial resolution $h$:

$$\rho_h(d_i) = 1 - \frac{6 \sum_{i=1}^{N_h} d_i(h)^2}{N_h(N_h^2 - 1)}$$

where $N_h$ is number of patches of radius $h$ and $d_i(h) = x_i(h) - y_i(h)$ is the difference between ranks of simulated population $x_i$ and population $y_i$ recorded in [7] of settlement $i$. We sweep $h \in [1, 100]$ km, sampling 1000 settlements at each resolution $h$ and compile the results on Fig. 4(a). $h$ can be compared to various reference scales such as the average distance to the nearest neighbor, average one-hour walking distance, average radius of a district, and average radius of a province at 1.4, 3, 25, and 80 km respectively. At each scale, our results and the data in [7] highly and statistically significantly correlated. For $h \geq 10$ km, the spatial distributions of the simulated and observed populations are almost identical. This result lends credence to our assumption that population distribution is driven by land availability is sufficient to replicate real-life settlement patterns in rural Afghanistan in silico.

![Fig. 3. Comparison of the joint distributions of farm wealth and size for the simulated rural population and the FAO surveyed population pooled for the whole country. The FAO surveyed 4669 rural households with the mean household size of 11.6. Simulated population of Afghanistan in 2000 consists of $1.45 \times 10^6$ rural households with the mean size of 10.5. The total simulated urban population of Afghanistan in 2000 was set to $4.5 \times 10^6$ persons.](image-url)
Spatial correlation is one measure of validity for our approach. As displayed on Fig. 4, absolute errors of district-level population counts\(^1\) provide another such measure. For this purpose, we rescaled our simulated population of \(18.2 \times 10^6\) individuals uniformly across Afghanistan to \(23.1 \times 10^6\) so the simulated population of Afghanistan equals that in [7]. We observe that the assumption of uniform population growth does not hold in districts that have received the largest numbers of assisted refugee returnees from Iran and Pakistan in 2002–2008 [15]. Accounting for the inflow of returnees to districts and applying uniform population growth during 2002–2008 to fill in the gap between simulated population and that in [7], the maximum absolute district-level error does not exceed 10000 persons per district with a median of 2400.

![Fig. 4. Comparison of 2008 population distribution in [7] and the 2002 simulated population. Assuming only uniform population growth, the top ten districts with largest absolute errors are highlighted in yellow and correspond to cities of Kandahar, Kabul, Jalalabad and Militarlam.](attachment:image)

5 Conclusions

In this paper we developed a protocol that uses local surveys and partial remote sensing data to recover country-level virtual rural populations that can be used to model licit and illicit economic activities in countries that are mired in irregular warfare. Our approach unifies scattered datasets into an ontology that is consistent with our target multiagent model. Using Afghanistan as a case, we showed that the assumption that the distribution of rural population is driven

---

\(^1\) Afghanistan had 329 districts before 2004 when a major reorganization increased the number of districts to 397. Since June 2005, the Afghan Ministry of Interior has recognized 398 districts in 34 provinces. We use the original 329 figure.
by land availability is sufficient to recreate rural population below the district level and validated our procedure by an independent dataset.

References